On the maximum size of minimal definitive quartet sets

Chris Dowden

LIX, École Polytechnique, 91128 Palaiseau Cedex, France

Abstract

In this paper, we investigate a problem concerning quartets, which are a particular type of tree on four leaves. Loosely speaking, a set of quartets is said to be 'definitive' if it completely encapsulates the structure of some larger tree, and 'minimal' if it contains no redundant information. Here, we address the question of how large a minimal definitive quartet set on n leaves can be, showing that the maximum size is at least $2n - 8$ for all $n \geq 4$. This is an enjoyable problem to work on, and we present a pretty construction, which employs symmetry.

Keywords: Quartet, Minimal definitive quartet set, Binary tree

1. Introduction

The motivation for this paper comes from the field of phylogenetics, which involves the study of the 'tree of life' depicting all living things, as popularised by Charles Darwin. In such a representation, existing species are drawn as leaves of the tree, while their ancestors are shown as interior vertices.

In practice, the overall evolutionary (or 'phylogenetic') tree is built up by piecing together various smaller items of information. For example, if species u and v both have wings and species w and x do not, then it is likely that u and v have a common ancestor that is not shared by w and x, and so the path from u to v on the tree of life should not intersect the path from w to x.

The objective of this paper is to present a new result on quartets, which are a type of graph often used when reconstructing evolutionary trees in this way. We start by providing some necessary definitions.

A *phylogenetic tree* is a tree with no vertices of degree 2 in which the leaves are labelled (distinctly) and the interior vertices are not. A phylogenetic tree is called *binary* if all interior vertices have degree exactly 3, and a *quartet* is defined to be a binary phylogenetic tree with precisely four leaves (note that such a tree is unique up to labelling). We use the notation $uv|wx$ to denote a quartet that is labelled as in Figure 1.

We say that a phylogenetic tree T *displays* the quartet $uv|wx$ if u, v, w and x are all leaves in T and the path from u to v does not intersect the path from

Email address: dowden@lix.polytechnique.fr (Chris Dowden)

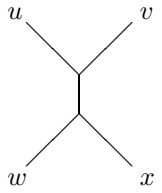

Figure 1: The quartet $uv|wx$.

w to x (or, equivalently, there exists a cut-edge in T which separates u and v from w and x). We say that T displays a set of quartets Q if T displays each individual quartet $q \in Q$.

For a set of quartets Q with total leaf-set $L(Q)$, we say that Q *defines* a tree T (or that Q is *definitive* for T) if T is the unique phylogenetic tree with leaf-set $L(Q)$ that displays Q. Note that many quartet sets will not define any tree, either because they contain quartets that are incompatible (e.g. $\{uv|wx, uw|vx\}$) or because they are not informative enough to be particular to one tree (e.g. $\{uv|wx, uv|wy\}$ is displayed by four different phylogenetic trees with leaf-set $\{u, v, w, x, y\}$, as shown in Figure 2). An example of a quartet

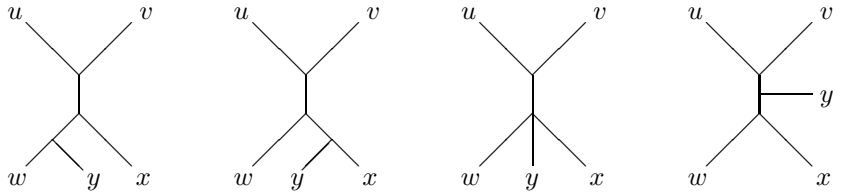

Figure 2: Four different phylogenetic trees displaying the quartets $uv|wx$ and $uv|wy$.

set that is definitive is $\{uv|wx, ux|wy\}$, which can be seen to define the left-most tree in Figure 2. Finally, we say that Q is a *minimal* definitive quartet set (or that Q is *minimally definitive*) if Q defines some tree T but, for all $q \in Q$, $Q - q$ does not define T (for example, $\{uv|wx, ux|wy\}$ is minimally definitive, but not $\{uv|wx, ux|wy, uv|wy\}$).

It is fairly straightforward to see that if Q defines T, then T must be binary and Q must *distinguish* every interior edge of T (a quartet $uv|wx$ is said to distinguish an edge $e \in T$ if e is the unique cut-edge in T that separates u and v from w and x). Thus, since a binary tree on n leaves always has exactly $n-3$ interior edges, it follows that every definitive quartet set on $\{1, 2, \ldots, n\}$ must contain at least $n - 3$ quartets. Furthermore, it is known that for any binary phylogenetic tree T on n leaves, there is indeed a set of $n-3$ quartets that does define T (see, for example, [2] Corollary 6.3.10).

Hence, the remaining interest in minimal definitive quartet sets lies in the question of how large they can be. Examples have been produced that have size greater than $n-3$, but until recently it was thought that the maximum possible size would be bounded by $n + c$ for some fixed constant c. However, Humphries

([1], Theorem 3.4.1) then proved that there actually exist examples with size at least $\frac{3}{2}n - 6$, for all $n \geq 4$. In this paper, we will improve matters still further by constructing minimal definitive quartet sets of size $2n - 8$, for all $n \geq 5$.

2. Main Section

This section will culminate in the inductive construction of minimal definitive quartet sets of size $2n - 8$. The structure of the section will be as follows: we shall start by stating three lemmas that will be useful to us; we shall then prove the result for $n = 6$, which will be the base case for our induction; we shall then also prove the $n = 7$ case, as a way to convey the ideas of the inductive step; and finally we shall prove the full result.

We start by making explicit a result that we have already noted:

Lemma 1 ([2], Proposition 6.8.4). *Let Q be a set of quartets that defines a tree T. Then each interior edge of T must be distinguished by at least one quartet in Q.*

The converse of Lemma 1 is known not to be true in general. However, the following result, which will play an extremely important role in our construction, comes close:

Lemma 2 ([2], Theorem 6.8.8). *Let Q be a set of quartets containing a common leaf, and let T be a tree displaying Q for which each interior edge is distinguished by at least one quartet in Q. Then Q defines T.*

Often, a set of quartets Q can be used to 'infer' a further quartet q, in the sense that every phylogenetic tree displaying Q must also display q. The notation $Q \vdash q$ is used to denote such inferences. Numerous examples are known, but we shall only use one very simple one:

Lemma 3. $\{ab|de, bc|de\} \vdash ac|de$.

As well as the three lemmas that we have stated, *caterpillar* trees will also play a major role in our proofs. The caterpillar tree on i leaves, which we shall denote by T_i, is defined via Figure 3.

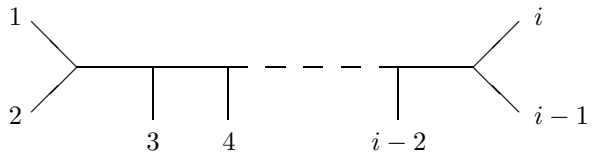

Figure 3: The caterpillar tree T_i.

We shall now give an example of a minimal definitive quartet set on six leaves that has size four, thus fulfilling our $2n-8$ target. Such sets have already been produced before now, but ours has a nice reversible symmetry to it that will later prove significant.

Lemma 4. *The set of quartets*

$$Q_6 = \{12|35, 13|46, 12|56, 24|56\}$$

is minimally definitive for the caterpillar tree T_6.

Proof Let us first show that Q_6 is definitive. Note that we have $\{12|56, 24|56\} \vdash 14|56$, by Lemma 3, and hence $Q_6 \vdash \{12|35, 13|46, 14|56\}$. Since the three quartets in this subset all contain the common leaf 1 and collectively distinguish each interior edge of T_6, definitiveness then follows automatically from Lemma 2.

It remains to show that Q_6 is *minimally* definitive. If not, then there exists a quartet $q \in Q_6$ such that $Q_6 - q$ defines T_6. By Lemma 1, the only possibility for q is $12|56$. However, the tree T' shown in Figure 4 displays $Q_6 - 12|56$, and

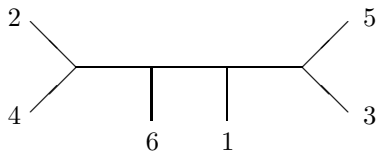

Figure 4: A tree T' displaying the quartet set $Q_6 - 12|56$.

T' is certainly distinct from T_6. Hence, it follows that Q_6 is indeed minimally definitive. □

We shall now see how to use the set Q_6 from Lemma 4 to produce a minimal definitive quartet set on seven leaves that has size six. This example, combined with the paragraph of discussion after the proof, is intended to help make clear the general strategy, which utilises the symmetry properties that we have observed, but the reader is free to proceed straight to the full proof of Theorem 6 if he so wishes.

Example 5. *The set of quartets*

$$Q_7 = \{12|35, 13|46, 12|57, 24|57, 13|67, 35|67\}$$

is minimally definitive for the caterpillar tree T_7.

Proof As a rigorous proof is implicitly included within that to Theorem 6, we shall provide a slightly more informal treatment here. Firstly, note that Table 1 shows that we can again use Lemma 2 to prove definitiveness (it is worth observing the way that the quartets are paired up here). By Lemma 1, it then only remains to provide suitable trees T'' and T''' displaying $Q_7 - 12|57$ and $Q_7 - 13|67$, respectively. But note that T'' (see Figure 5) can easily be formed from the tree T' in Figure 4 (it is important to note the similarity between Q_6 and Q_7), while the symmetry of Q_7 allows us to take T''' to be the same as T'', but with the numbers reversed! □

$$\left.\begin{array}{l} 12|35 \\ 13|46 \\ 12|57 \\ 24|57 \end{array}\right\} \vdash 14|57$$
$$\left.\begin{array}{l} 13|67 \\ 35|67 \end{array}\right\} \vdash 15|67$$

Table 1: Some inferences that can be made from the quartet set Q_7.

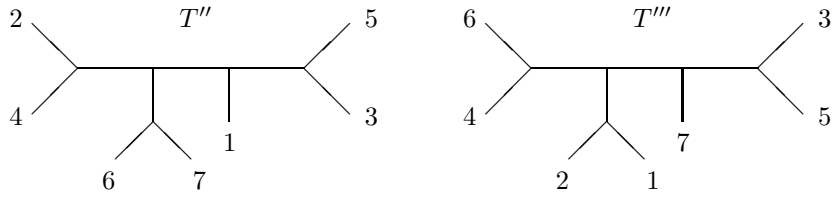

Figure 5: Trees T'' and T''' displaying the quartets $Q_7 - 12|57$ and $Q_7 - 13|67$, respectively.

As we shall now see, the inductive step in the full proof follows exactly the same procedure as in the example above. We shall always use caterpillars, and we shall show definitiveness by always adding an extra pair of quartets that together infer $1(n-2)|(n-1)n$. Proving minimality will then come down to constructing various trees of the form $Q - q$. All but one of these will be formed from the trees of the previous stage of the induction, while the additional tree will be created by reversing numbers.

Theorem 6. *Let $n \geq 5$ be some positive integer. Then there exists a minimal definitive quartet set on n leaves that has size $2n - 8$.*

Proof We have already noted the result for $n \in \{5, 6\}$ (and also for $n = 7$, for those that have read Example 5), so we shall now proceed inductively, using the set Q_6 defined in Lemma 4 as our base. Let us use $q_{6,1}$, $q_{6,2}$, $q_{6,3}$ and $q_{6,4}$, respectively, to denote the quartets 12|35, 13|46, 12|56 and 24|56, in that order (so $Q_6 = \{q_{6,1}, q_{6,2}, q_{6,3}, q_{6,4}\}$). For $k \geq 7$, let us then define $Q_k = \{q_{k,1}, q_{k,2}, \ldots, q_{k,2k-8}\}$ recursively from Q_{k-1} as follows: (i) for all $i \leq 2k-12$, set $q_{k,i} = q_{k-1,i}$; (ii) for $i \in \{2k-11, 2k-10\}$, set $q_{k,i}$ to be the same as $q_{k-1,i}$, but with occurrences of $k-1$ replaced by k; and (iii) set $q_{k,2k-9} = 1(k-4)|(k-1)k$ and $q_{k,2k-8} = (k-4)(k-2)|(k-1)k$. It can be checked that this procedure produces the set Q_7 defined in Example 5.

Note that $|Q_k| = 2k-8$ and so, by induction, it now suffices to prove that Q_k is minimally definitive for the caterpillar tree T_k given that Q_{k-1} is minimally definitive for T_{k-1}. This is precisely what we shall now do.

First, let us check that T_k displays Q_k. Note that T_k displays all quartets $uv|wx$ for $u < v < w < x$, and $q_{k,2k-9}$ and $q_{k,2k-8}$ are certainly of this form. By induction, we can see that all other quartets in Q_k also satisfy this property, and so T_k does indeed display Q_k.

Next, we shall show that Q_k is definitive for T_k, using the same argument as for when $k = 6$. By induction, we may assume that $\{q_{k,1}, q_{k,2}, \ldots, q_{k,2k-12}\} \vdash$

$\{12|35, 13|46, 14|57, \ldots, 1(k-4)|(k-3)(k-1)\}$. Note $q_{k,2k-11} = 1(k-5)|(k-2)k$ and $q_{k,2k-10} = (k-5)(k-3)|(k-2)k$, so $\{q_{k,2k-11}, q_{k,2k-10}\} \vdash 1(k-3)|(k-2)k$ by Lemma 3 (and so the induction does hold). Finally, Lemma 3 also implies that $\{q_{k,2k-9}, q_{k,2k-8}\} \vdash 1(k-2)|(k-1)k$. Hence, just as with the case when $k = 6$, we may use Lemma 2 to deduce that Q_k is definitive for T_k.

It now only remains for us to show that Q_k is *minimally* definitive. To do this, we need to show that for each $q_{k,i}$ there exists a tree $T_{k,i} \neq T_k$ that displays $Q_k - q_{k,i}$.

For $i \leq 2k-10$, we can take $T_{k,i}$ to be the tree formed from $T_{k-1,i}$ by replacing vertex $k-1$ with a 'cherry' $\{k-1, k\}$ (by which we mean a rooted binary tree with leaf-set $\{k-1, k\}$ — for example, the tree T'' in Figure 5 is formed from the tree T' in Figure 4 by replacing vertex 6 with the cherry $\{6, 7\}$). It is clear that $T_{k,i} \neq T_k$, since $T_{k-1,i} \neq T_{k-1}$. The proof that $T_{k,i}$ displays $Q_k - q_{k,i}$ follows from observing that (a) $T_{k,i}$ displays all quartets that are displayed by $T_{k-1,i}$, since $T_{k-1,i}$ is a subgraph of $T_{k,i}$, (b) for $w < k-1$, $T_{k,i}$ displays the quartet $uv|wk$ if it displays $uv|w(k-1)$ (since $\{k-1, k\}$ forms a cherry), and hence $T_{k,i}$ displays $uv|wk$ if $T_{k-1,i}$ displays $uv|w(k-1)$, and (c) $T_{k,i}$ displays all quartets of the form $uv|(k-1)k$, again using the fact that $\{k-1, k\}$ forms a cherry.

For $i = 2k-8$, we may appeal to Lemma 1, and so this only leaves the case $i = 2k-9$, for which we have the quartet $q_{k,2k-9} = 1(k-4)|(k-1)k$. To deal with this, let us take $T_{k,2k-9}$ to be the tree formed from $T_{k,3}$ (which displays $Q_k - 12|57$) by 'reversing' all the numbers, i.e. 1 becomes k, 2 becomes $k-1$, 3 becomes $k-2$, and so on. Note that $T_{k,2k-9} \neq T_k$, since T_k is the 'reverse' of itself, and so it only remains to show that $T_{k,2k-9}$ displays $Q_k - q_{k,2k-9}$, which we shall now do.

First, note that the tree $T_{k,3}$ was formed by taking a tree displaying $Q_6 - 12|56$, replacing vertex 6 with a cherry $\{6, 7\}$, then replacing vertex 7 with a cherry $\{7, 8\}$, replacing vertex 8 with a cherry $\{8, 9\}$, and so on until replacing vertex $k-1$ with a cherry $\{k-1, k\}$. Hence, $T_{k,3}$ must display every quartet of the form $uv|wx$ for $u < v < w < x$ and $w \geq 6$, and so $T_{k,2k-9}$ must display every quartet of the form $ab|cd$ for $a < b < c < d$ and $b \leq k-5$.

This immediately covers every quartet in $Q_k - q_{k,2k-9}$ apart from three: $q_{k,2k-12} = (k-6)(k-4)|(k-3)(k-1)$, $q_{k,2k-10} = (k-5)(k-3)|(k-2)k$ and $q_{k,2k-8} = (k-4)(k-2)|(k-1)k$. Furthermore, since these are the 'opposites' of $q_{k,4} = 24|57$, $q_{k,2} = 13|46$ and $q_{k,1} = 12|35$, which are all displayed by $T_{k,3}$, we find that these three remaining quartets are also all displayed by $T_{k,2k-9}$. Hence, $T_{k,2k-9}$ displays $Q_k - q_{k,2k-9}$, and so we are done. □

3. Questions

The obvious question is whether $2n-8$ can be improved upon, and it would be interesting to know of any better examples. Throughout this paper, we have only used caterpillar trees, partly for simplicity, and so another question of interest would be to ask whether caterpillars can always be relied upon to

provide the extremal cases. Finally, it would also be nice to obtain some sort of upper bound on the maximum possible size of a minimal definitive quartet set, other than the trivial $\binom{n}{4}$.

Acknowledgements

I would like to thank Charles Semple for introducing me to the problem, and the reviewers for their helpful comments.

References

References

[1] P. J. Humphries, Combinatorial aspects of leaf-labelled trees (PhD thesis, 2008), available at http://hdl.handle.net/10092/1801.

[2] C. Semple, M. Steel, Phylogenetics, Oxford University Press, Oxford, 2003.